Estimation with Selected Binomial Information

or

Do You Really Believe that Dave Winfield is Batting .471?

George Casella[1]
Cornell University, Ithaca, NY

Roger L. Berger
North Carolina State University, Raleigh, NC

BU-1197-M April 1993

AMS 1990 Subject Classification: Primary 62F99; Secondary 62E99.

Key words and phrases: Selection bias, EM algorithm, Gibbs sampling, Brownian motion.

[1] Research supported by National Science Foundation Grant No. DMS9100839 and National Security Agency Grant No. 90F-073.

Abstract

Often sports announcers, particularly in baseball, provide the listener with exaggerated information concerning a player's performance. For example, we may be told that Dave Winfield, a popular baseball player, has hit safely in 8-of-his-last-17 chances (.471). This is biased, or selected information, as the "17" was chosen to maximize the reported percentage. We model this as observing a maximum success rate of a Bernoulli process, and show how to construct the likelihood function for a player's true batting ability. The likelihood function is a high degree polynomial, but can be computed exactly. Alternatively, the problem yields to solutions based on either the EM algorithm or Gibbs sampling. Using these techniques, we compute maximum likelihood estimators, Bayes estimators, and associated measures of error. We also show how to approximate the likelihood using a Brownian motion calculation. We find that, although it is difficult to construct good estimators from selected information, we seem to be able to estimate better than expected, particularly when using prior information. The estimators are illustrated with data from the 1992 Major League baseball season.

1. Introduction

Sports announcers, in particular baseball announcers, often use hyperbolical descriptions of a player's ability. For example, when Dave Winfield, a popular baseball player, is batting, rather than report his current batting average (number of hits divided by number of at-bats), it might be said, "He is really hitting well these days. He is 8 for his last 17." This is clearly selectively reported data, biased upward from the player's actual average. However, with models that take this bias into account, we should be able to use the selectively reported data to recover an estimate of the player's true ability. In this article we explore various methodologies for doing this.

1.1. Background

Research in estimation and modeling from selectively reported data has always been of interest, and has many applications other than analyzing baseball data. We will not attempt a thorough literature review here, but will only describe some general directions that such research has taken. Perhaps the most widespread use of selection-bias methodology is in the area of meta-analysis. Starting from work of Rosenthal (1979), researchers have worried about the effect of selectively reported data when combining results of different studies, where the selection is mainly through publication bias (only publishing significant studies). These concerns are summarized and reviewed by Iyengar and Greenhouse (1988) and, more recently, in a trio of papers in *Statistical Science* (Mosteller and Chalmers, 1992; Dear and Begg, 1992; Hedges, 1992). Cleary (1993) has used these selection models, along with likelihood theory and Gibbs sampling, to construct estimates of effects based on publication-biased data.

A Bayesian approach to inference from selected data was taken by Bayarri and DeGroot (1986a, b, 1991). A major lesson to be learned from their work is that the uncorrected MLE can be exceedingly bad. In our baseball data it is quite obvious that the naive estimate of Winfield's batting ability, $8/17 = .471$, is vastly incorrect. In other, more complicated, situations this might not be so obvious.

Perhaps the methodology that is most similar to what is done here is that of Dawid and Dickey (1977). They were concerned with the influence of selectively reported data on the likelihood function, and how such influence can be accounted for. In particular, they considered an example where the selectively reported data is the maximum of sums of Bernoulli random variables, an example that is closely related to our situation. More recently, Carlin and Gelfand (1992) have studied parametric likelihood inference in "record-breaking" data, that is, data that are a series of records, or maxima. They discuss many applications of their models, including sporting events, meteorology, and industrial stress testing. In particular, they model an underlying regression that attempts to explain the increasing sequence of means, and illustrate their techniques using data on Olympic record high jumps. The selected data that we are concerned with here may be thought of as

a special case of "record breaking" data. However, our models and estimation methodologies are different from previous approaches.

1.2. Information

To make inferences from our selected data, we must make some assumptions about the data we see. For example, when we are told that Dave Winfield is 8-for-his-last-17, we assume that the "17" is chosen because the ratio 8/17 is the maximum ratio of hits-to-at-bats. There is some hidden information in this number. For example, we know that on his 18^{th} previous at-bat he didn't get a hit, otherwise the announcer would have reported $9/18 > 8/17$. Moreover, our naive estimate of his ability should not be $8/17 = .471$, but $8/18 = .444$, since we know that the 18^{th} previous at-bat was a failure. More precisely, we assume that a baseball player's sequence of at-bats is a sequence of Bernoulli(θ) random variables, $X_1, \ldots, X_n$, where the X_is are in chronological order. On the n^{th} at-bat, the player's batting average is $\hat{p} = \sum_{i=1}^{n} X_i \Big/ n$. This is the maximum likelihood estimator (MLE) based on observing the entire (complete) data set. However, we assume that the data reported is k^* hits out of the last m^* at-bats, where k^* and m^* satisfy

$$r^* = \frac{k^*}{m^*} \geq \max_{m^* \leq i < n} \frac{X_n + X_{n-1} + \cdots + X_{n-i}}{i+1} \tag{1.1}$$

Note that the quantity $\left(X_n + X_{n-1} + \cdots + X_{n-i}\right) \Big/ (i+1)$ is just a player's batting average in the previous $i+1$ at-bats. Thus, we are assuming only that there is no higher hits-to-at-bats ratio in the unreported data than the reported ratio of $r^* = k^*/m^*$. There may be a higher ratio in the last m^* at-bats, for example, perhaps the batter was 1-for-1 in his last at-bat. In practice, with the exception of similar trivial cases, r^* will usually represent the maximum ratio of hits to at-bats. Also notice that the exact mechamism of choosing m^* need not be known, we only need assume that (1.1) is satisfied.

1.3. Data

We will illustrate the selected data information on data from the 1992 Major League Baseball season. Our first data set is the record of all of Dave Winfield's 1992 at-bats, and whether the at-bat resulted in a hit or out. (Winfield actually made 670 plate appearances in 1992, but 87 of these were not "at-bats", since they resulted in either a walk or a sacrifice. Thus, there were 583 at-bats.)

The data for Dave Winfield are displayed in Figure 1. The dashed line represents his batting average (ratio of hits to at-bats) for each at-bat. It can be seen that this value settles down quickly, and remains close to $169/583 = .290$, Winfield's final batting average for the 1992 season. For each at-bat, this ratio is also the maximum likelihood estimator given that we have observed the entire sequence of all previous at-bats or, equivalently, that we know the total number of hits up to the given at-bat.

The solid line in Figure 1 is a running sequence of values of r^*. For each at-bat, this number is the maximum ratio of hits to at-bats, counting backwards in time from the given at-bat. In the calculation of r^* we required $m^* > 10$, which merely serves to eliminate trivial cases (1 for his last 1), and smooths out the picture somewhat (eliminating multiple peaks at 2-for-3, 4-for-7, etc.). Thus, Figure 1 shows 573 at-bats, starting from at-bat number 11.

Lastly, the dotted line in Figure 1 is a running plot of the selected data MLE, the maximum likelihood estimate of Winfield's batting ability based on only observing k^*, m^*, and n = at-bat number. This estimator is one of the main objects of investigation in this paper, and will be discussed in detail in later sections.

We will also analyze a similar data set composed of the 1992 won-loss record of the New York Mets, which is pictured in Figure 2. The Mets played 162 games and, as before, we show the 152 games from game 11 to game 162. The values of r^* (solid line) are somewhat less variable than Winfield's, and again the complete data MLE (dashed line) quickly settles down to the final winning percentage of the Mets, $72/162 = .444$. The selected data MLE, however, still remains quite variable.

1.4. Summary

In Section 2 we show how to calculate the exact likelihood based on observing the selected information k^*, m^* and n. We do the calculations two ways, one that uses an exact combinatoric derivation of the likelihood function, and one based on a Gibbs sampling algorithm. The combinatorial derivation, and the resulting likelihood, are quite complicated. However, an easily implementable (albeit computer-intensive) Gibbs algorithm yields likelihood functions that are virtually identical. In Section 3 we consider maximum likelihood point estimation and estimation of standard errors, and also show how to implement Bayesian estimation via the Gibbs sampler. We also calculate maximum likelihood point estimates in a number of ways, using the combinatorial likelihood, the EM algorithm, and a Gibbs sampling-based approximation, and show how to estimate standard errors for these point estimates.

Section 4 adapts methodology from sequential analysis to derive a Brownian motion-based approximation to the likelihood. The approximation also yields remarkably accurate MLE values. Section 5 is a discussion that relates all of this methodology back to the baseball data that originally suggested it. Finally, there is an Appendix with some technical details.

2. Likelihood Calculations

In this section we show how to calculate the likelihood function exactly. We use two methods, one based on a combinatorial derivation, and one based on Markov chain methods.

Recall that the data, $X_1,\ldots,X_n$ are in chronological order. However, for selectively reported data like we are considering, it is easier to think in terms of the reversed sequence, looking backward from time n, so we now redefine the data in terms of the reversed sequence. Also, we want to distinguish between the reported and unreported data. We define $\mathbf{Y} = (Y_1,\ldots,Y_{m^*+1})$ by $Y_i = X_{n-i+1}$ and $\mathbf{Z} = (Z_1,\ldots,Z_m)$ by $Z_i = X_{n-m^*-i}$, where $m = n-(m^*+1)$. Thus, $\mathbf{Y}$ is the reported data (with Y_1 being the most recent at-bat, Y_2 being the next most recent, etc.), including $Y_{m^*+1} = X_{n-m^*}$, which we know to be 0. There are k^* 1s in $\mathbf{Y}$. $\mathbf{Z}$ is the unreported data. We know that the vector $\mathbf{Z}$ satisfies

$$\frac{k^* + \sum_{i=1}^{j} Z_i}{m^*+1+j} \le r^* = \frac{k^*}{m^*} \quad \text{for all} \quad j = 1,\ldots,m\,. \tag{2.1}$$

This is assumption (1.1), that there is no higher ratio than the reported r^* in the unreported data.

The likelihood, given the reported data $\mathbf{Y} = \mathbf{y}$, will be denoted by $L(\theta\,|\,\mathbf{y})$. (Generally, random variables will be denoted by upper case letters, and their observed values by the lower case counterparts.) It is proportional to

$$L(\theta\,|\,\mathbf{y}) \;\propto\; \theta^{k^*}(1-\theta)^{m^*+1-k^*} \sum_{\mathbf{z}\in\mathcal{Z}^*} \theta^{S_{\mathbf{z}}}(1-\theta)^{m-S_{\mathbf{z}}}\,, \tag{2.2}$$

where $\mathcal{Z}^*$ is the set of all vectors $\mathbf{z} = (z_1,\ldots,z_m)$ that never give a higher ratio than r^* [see (2.1)], and $S_{\mathbf{z}} = \sum_{i=1}^{m} z_i$ = number of 1s in the unreported data. Dawid and Dickey (1977) call the factor $\theta^{k^*}(1-\theta)^{m^*-k^*}$, the face-value likelihood, and the remainder of the expression the correction factor. The correction factor is the correction to the likelihood that is necessary since the data was selectively reported.

2.1. Combinatorial Calculations

An exact expression for the sum over $\mathcal{Z}^*$ can be given in terms of constants i^*, $n_1,\ldots,n_{i^*}$, and $c_1,\ldots,c_{i^*}$, which we now define. Let i^* = largest integer that is less than $(n-m^*)(1-r^*)$, and define

$$n_i = \left[\frac{i}{1-r^*}\right], \qquad i = 1,\ldots,i^*\,,$$

where [a] = greatest integer less than or equal to a. Now define constants c_i recursively by $c_1 = 1$ and

$$c_i = \binom{n_i-1}{i-1} - \sum_{j=1}^{i-1} c_j \binom{n_i-1-n_j}{i-j}, \qquad i = 2,\ldots,i^*\,.$$

Then (2.2) can be written

$$L(\theta\,|\,\mathbf{y}) \;\propto\; \theta^{k^*}(1-\theta)^{m^*+1-k^*}\left\{1 - \sum_{i=1}^{i^*} c_i\,\theta^{n_i+1-i}(1-\theta)^{i-1}\right\}. \tag{2.3}$$

The equivalence of (2.2) and (2.3) is proved in Appendix 1. If $i^* = 0$, which will be true if r^* is large and $n-m^*$ is small, the sum in (2.3) is not present and the likelihood is just

$$L(\theta \mid \mathbf{y}) \propto \theta^{k^*}(1-\theta)^{m^*+1-k^*} .$$

The constants c_i grow very rapidly as i increases. So if i^* is even moderately large, care must be taken in their computation. They can be computed exactly with a symbolic processor, but this can be time consuming. So we now look at alternate ways of computing the likelihood, and in Section 4 we consider an approximation of $L(\theta \mid \mathbf{y})$.

2.2. Sampling-Based Calculations

As an alternative to the combinatorial approach to calculating $L(\theta \mid \mathbf{y})$, we can implement a sampling-based approach using the Gibbs sampler. We can interpret equation (2.2) as stating

$$L(\theta \mid \mathbf{y}) = \sum_{\mathcal{Z}^*} L(\theta \mid \mathbf{y}, \mathbf{z}) \tag{2.4}$$

where $\mathbf{z} = (z_1, \ldots, z_m)$ are the unobserved Bernoulli outcomes, $\mathcal{Z}^*$ is the set of all such possible vectors, and $L(\theta \mid \mathbf{y}, \mathbf{z})$ is the likelihood based on the complete data. Equation (2.4) bears a striking resemblance to the assumed relationship between the "complete data" and "incomplete data" likelihoods for implementation of the EM algorithm and, in fact, can be used in that way. We will later see how to implement an EM algorithm, but first we show how to use (2.4) to calculate $L(\theta \mid \mathbf{y})$ using the Gibbs sampler.

We assume that $L(\theta \mid \mathbf{y})$ can be normalized in θ, that is, $\int_\Theta L(\theta \mid \mathbf{y})\, d\theta < \infty$. (This is really not a very restrictive assumption, as most likelihoods will have finite integrals.) Denote the normalized likelihood by $L^*(\theta \mid \mathbf{y})$, so

$$L^*(\theta \mid \mathbf{y}) = \frac{L(\theta \mid \mathbf{y})}{\int_\Theta L(\theta \mid \mathbf{y})\, d\theta} . \tag{2.5}$$

Since $L(\theta \mid \mathbf{y})$ can be normalized, so can $L(\theta \mid \mathbf{y}, \mathbf{z})$ of (2.4). Denoting that normalized likelihood by $L^*(\theta \mid \mathbf{y}, \mathbf{z})$, we now can consider both $L^*(\theta \mid \mathbf{y})$ and $L^*(\theta \mid \mathbf{y}, \mathbf{z})$ as density functions in θ. Lastly, from the unnormalized likelihoods, we define

$$k(\mathbf{z} \mid \mathbf{y}, \theta) = \frac{L(\theta \mid \mathbf{y}, \mathbf{z})}{L(\theta \mid \mathbf{y})} , \tag{2.6}$$

an equation reminiscent of the EM algorithm. The function $k(\mathbf{z} \mid \mathbf{y}, \theta)$ is a density function, and defines the density of $\mathbf{Z}$ conditional on $\mathbf{y}$ and θ. If we think of $\mathbf{z}$ as the "missing data" or, equivalently, $\mathbf{y}$ as the "incomplete data" and $(\mathbf{y}, \mathbf{z})$ as the "complete data", we can use the unnormalized likelihoods in a straightforward implementation of the EM algorithm. However, we have more. If we iteratively sample between $k(\mathbf{z} \mid \mathbf{y}, \theta)$ and $L^*(\theta \mid \mathbf{y}, \mathbf{z})$, that is, sample a sequence $\mathbf{z}_1$,

θ_1, $\mathbf{z}_2$, θ_2, $\mathbf{z}_3$, θ_3,..., then we can approximate the actual normalized likelihood by

$$\hat{L}^*(\theta \mid \mathbf{y}) \approx \frac{1}{M} \sum_{i=1}^{M} L^*(\theta \mid \mathbf{y}, \mathbf{z}_i) \tag{2.7}$$

with the approximation improving as $M \to \infty$ (see the Appendix for details). Thus we have a sampling-based exact calculation of the true likelihood function.

Note that this sampling-based strategy for calculating the likelihood differs from some other strategies that have been used. The techniques of Geyer and Thompson (1992) for calculating likelihoods, are based on a different type of Monte Carlo calculation, one that is not based on Gibbs sampling. That is also the technique employed by Carlin and Gelfand (1992) and Gelfand and Carlin (1991). The technique used here, which closely parallels the implementation of the EM algorithm, is discussed (but not implemented) by Smith and Roberts (1993).

Implementing equation (2.7) is quite easy. The likelihood $L^*(\theta \mid \mathbf{y}, \mathbf{z})$ is the normalized complete data likelihood, so

$$L^*(\theta \mid \mathbf{y}, \mathbf{z}) = \frac{\Gamma(n+2)}{\Gamma(k^* + S_{\mathbf{z}} + 1)\,\Gamma(n - k^* - S_{\mathbf{z}} + 1)} \theta^{k^* + S_{\mathbf{z}}} (1-\theta)^{n - k^* - S_{\mathbf{z}}} \tag{2.8}$$

(recall $S_{\mathbf{z}} = \sum_{i=1}^{m} z_i$ and $m = n - m^* - 1$). Thus, to calculate $\hat{L}^*(\theta \mid \mathbf{y})$ we use the following algorithm:

$$\begin{aligned} &0.\ \text{Initialize } \theta = \theta_0 . \\ &\text{For } j = 1, \ldots, M \\ &\quad 1.\ \text{Generate } \mathbf{z}_j \sim k(\mathbf{z} \mid \mathbf{y}, \theta_{j-1}) \\ &\quad 2.\ \text{Generate } \theta_j \sim L^*(\theta \mid \mathbf{y}, \mathbf{z}_j) . \end{aligned} \tag{2.9}$$

Since $L^*(\theta \mid \mathbf{y}, \mathbf{z})$ is a beta distribution with parameters $k^* + S_{\mathbf{z}} + 1$ and $n - k^* - S_{\mathbf{z}} + 1$, it is easy to generate the θs. To generate the $\mathbf{z}$s, from $k(\mathbf{z} \mid \mathbf{y}, \theta)$ the following simple rejection algorithm runs very quickly:

$$\begin{aligned} &1.\ \text{Generate } \mathbf{z} = (z_1, \ldots, z_m),\ z_i \text{ iid Bernoulli}(\theta) \\ &2.\ \text{Calculate } S_i = \left(k^* + \sum_{j=1}^{i} z_j \right) \Big/ (m^* + 1 + i), \qquad i = 1, \ldots, m \\ &3.\ \text{If } S_i \le r^* \text{ for every } i = 1, \ldots, m \text{ accept } \mathbf{z}, \text{ otherwise reject } \mathbf{z}. \end{aligned} \tag{2.10}$$

Implementing this algorithm using a 486DX2 computer with the Gauss programming language is very simple, and the running time is often quite short. The running time was increased only in situations where $n \gg m^*$. Then many constrained Bernoulli sequences were needed.

Figure 3 illustrates the Gibbs-sampled likelihoods for M = 1000, for $r^* = 8/17$ with a variety of values of n. As can be seen, the modes and variances decrease as n increases. If we plot the likelihoods calculated from the combinatorial formula (2.3), the differences are imperceptible.

3. Estimation

One goal of this article is to assess our ability to recover a good estimate of θ from the selectively reported data. We would be quite happy if our point estimate from the selected data is close to the MLE of the unselected data. However, as we shall see, this is generally not the case. Although, in some cases, we can do reasonably well, only estimation with strong prior information will do well consistently.

3.1. Exact Maximum Likelihood Estimation

Based on the exact likelihood of (2.3), we can calculate the MLE by finding the zeros of the derivative of $L(\theta \mid \mathbf{y})$. The likelihood is a high degree polynomial in θ, but symbolic manipulation programs can compute the constants and symbolically differentiate the polynomial. However, the zeros must be solved for numerically as the resulting expressions are too involved for analytical evaluation. In all the examples we have calculated, $L(\theta \mid \mathbf{y})$ is a unimodal function for $0 \leq \theta \leq 1$ and no difficulties were experienced in numerically finding the root.

We have calculated the MLE for several different data sets and the results are in Table 1. Results for four values of m^* and r^* and five values of m are given in the table. For each value of m^*, r^* and m, two values are given. The first is the exact MLE, computed by the method just described. The second is an approximate MLE that will be discussed in Section 4. Just consider the exact values for now.

⟨Table 1 about here⟩

The exact MLEs exhibit certain patterns that would be expected for this data.

1. The MLE never exceeds the naive estimate $k^*/(m^*+1) = m^*r^*/(m^*+1)$.
2. For fixed reported data m^* and r^*, the MLE decreases as m, the amount of unreported data, increases. It appears to approach a nonzero limit as m grows. Knowing that the ratio does not exceed r^* in a long sequence of unreported data should lead to a smaller estimate of θ than knowing only that the ratio does not exceed r^* in a short sequence.
3. For fixed r^* and m, the MLE increases to r^* as m^*, the amount of reported data increases.

This method of finding the MLE requires a symbolic manipulation program to calculate the constants c_i or else some careful programming to deal with large factorials. Also, the method can be slow if m is large. The values for $m = 200$ in Table 1 each took several minutes to calculate. Thus we are led to investigate other methods of evaluating the MLE, methods that do not use direct calculation of $L(\theta \mid \mathbf{y})$. Although these other methods are computationally intensive, they avoid the problem of dealing with the complicated exact likelihood function.

3.2. The EM Algorithm

As in Section 2.2, the incomplete data interpretation of the likelihood function allows for easy

implementation of the EM algorithm. With $\mathbf{y}$ = incomplete data and $(\mathbf{y},\mathbf{z})$ = complete data, we compute an EM sequence $\hat{\theta}_1, \hat{\theta}_2, \cdots$ by

$$\hat{\theta}_{i+1} = \frac{k^* + E(S_{\mathbf{z}} | \hat{\theta}_i)}{n}, \tag{3.1}$$

where $E(S_{\mathbf{z}} | \hat{\theta}_i)$ is the expected number of successes in the missing data. (The E-step and the M-step are combined into one step in (3.1).) More precisely, $S_{\mathbf{z}} = \sum_{j=1}^{m} Z_j$, where the Z_j are iid Bernoulli($\hat{\theta}_i$) and the partial sums satisfy the restrictions in (2.10). Such an expected value is virtually impossible to calculate analytically, but is quite easy to approximate using Monte Carlo methods. The resulting sequence $\hat{\theta}_1, \hat{\theta}_2, \ldots$ converges to the exact complete data MLE. In all of our calculations, the value of the EM-calculated MLE is indistinguishable from the MLE resulting from (2.3).

3.3. Approximate MLEs from the Gibbs Sampler

Equation (2.7), which relates the exact likelihood to the average of the complete data likelihoods, forms a basis for a simple approximation scheme for the MLE. Although "the average of the maxima is not the maximum of the average", we can use a Taylor series approximation to estimate the incomplete data MLE as a weighted average of the complete data MLEs.

An obvious approach is to expand each complete data likelihood in (2.7) around its MLE, $\hat{\theta}_i$, to get

$$\begin{aligned} L^*(\theta | \mathbf{y}, \mathbf{z}_i) &\approx L^*(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i) + (\theta - \hat{\theta}_i) L^{*'}(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i) + \frac{(\theta - \hat{\theta}_i)^2}{2} L^{*''}(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i) \\ &= L^*(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i) + \frac{(\theta - \hat{\theta}_i)^2}{2} L^{*''}(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i), \end{aligned} \tag{3.2}$$

since $L^{*'}(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i) = 0$. Now substituting into (2.7) yields

$$\hat{L}^*(\theta | \mathbf{y}) \approx \frac{1}{M} \sum_{i=1}^{M} \left[L^*(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i) + \frac{(\theta - \hat{\theta}_i)^2}{2} L^{*''}(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i) \right]$$

and differentiating with respect to θ yields the approximate MLE

$$\hat{\theta}_A = \frac{\sum_{i=1}^{M} \hat{\theta}_i \, L^{*''}(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i)}{\sum_{i=1}^{M} L^{*''}(\hat{\theta}_i | \mathbf{y}, \mathbf{z}_i)}.$$

It turns out, however, that this approximation is not very accurate. A possible reason for this is the oversimplification of (3.2), which ignores most of the computed information. In particular, for $j \neq i$, the information in $\hat{\theta}_j$ is not used when expanding $L^*(\theta | \mathbf{y}, \mathbf{z}_i)$. Thus, we modify (3.2) into a "double"

Taylor approximation. We first calculate an average approximation for each $L^*(\theta \mid \mathbf{y}, \mathbf{z}_i)$, averaging over all $\hat{\theta}_j$, and then average over all $L(\theta \mid \mathbf{y}, \mathbf{z}_i)$. We now approximate the incomplete data likelihood $L^*(\theta \mid \mathbf{y})$ with

$$L^*(\theta \mid \mathbf{y}) = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \left[L^*(\hat{\theta}_j \mid \mathbf{y}, \mathbf{z}_i) + (\theta - \hat{\theta}_j) L^{*'}(\hat{\theta}_j \mid \mathbf{y}, \mathbf{z}_i) + \frac{(\theta - \hat{\theta}_j)^2}{2} L^{*''}(\hat{\theta}_j \mid \mathbf{y}, \mathbf{z}_i) \right].$$

Differentiating yields the approximate MLE

$$\hat{\hat{\theta}}_A = \frac{\sum_{i,j} \hat{\theta}_j L^{*''}(\hat{\theta}_j \mid \mathbf{y}, \mathbf{z}_i) - \sum_{i,j} L^{*'}(\hat{\theta}_j \mid \mathbf{y}, \mathbf{z}_i)}{\sum_{i,j} L^{*''}(\hat{\theta}_j \mid \mathbf{y}, \mathbf{z}_i)} . \tag{3.3}$$

From (2.8), denoting the constant by $C(\mathbf{z}_i)$, we have

$$L^{*''}(\hat{\theta}_i \mid \mathbf{y}, \mathbf{z}_i) = -nC(\mathbf{z}_i)\, \hat{\theta}_i^{k^* + S_z - 1} (1 - \hat{\theta}_i)^{n - k^* - S_z - 1} , \tag{3.4}$$

and substituting this into (3.3) gives our approximate MLE. Table 2 compares the approximate MLE of (3.3) to the exact value found by differentiating the exact likelihood of (2.3). It can be seen that the Gibbs approximation is reasonable, but certainly not as accurate as the EM calculation.

⟨Table 2 about here⟩

3.4. Bayes Estimation

It is relatively easy to incorporate prior information into our estimation techniques, especially when using the Gibbs sampling methodology of Section 2.2. More importantly, in Major League Baseball there is a wealth of prior information. For any given player or team, the past record is readily available in sources such as the Baseball Encyclopedia (1988).

If we assume that there is prior information available in the form of a beta(a, b) distribution then, analogous to Section 2.2, we have the two full conditional posterior distributions

$$\pi(\theta \mid \mathbf{y}, \mathbf{z}, a, b) = \frac{\Gamma(n + a + b)}{\Gamma(k^* + S_z + a)\, \Gamma(n - k^* - S_z + b)} \theta^{k^* + S_z + a - 1} (1 - \theta)^{n - k^* - S_z + b - 1} \tag{3.5}$$

$k(\mathbf{z} \mid \mathbf{y}, \theta, a, b)$ = as in (2.6) and (2.10), with the Bernoulli parameter = θ .

Running a Gibbs sampler on (3.5) is straightforward, and the posterior distribution of interest is given by

$$\hat{\pi}(\theta \mid \mathbf{y}, a, b) = \frac{1}{M} \sum_{i=1}^{M} \pi(\theta \mid \mathbf{y}, \mathbf{z}_i, a, b) . \tag{3.6}$$

For point estimation, we usually use the posterior mean, given by

$$\begin{aligned}\hat{E}(\theta \mid \mathbf{y}, a, b) &= \frac{1}{M} \sum_{i=1}^{M} E(\theta \mid \mathbf{y}, \mathbf{z}_i, a, b) \\ &= \frac{1}{M} \sum_{i=1}^{M} \frac{k^* + S_{\mathbf{z}_i} + a}{n + a + b}\end{aligned} \tag{3.7}$$

By using a beta(1,1) prior we get the likelihood function as the posterior distribution. Table 2 also shows the values of this point estimator, and we see that it is a very reasonable estimate.

Since the available prior information in baseball is so good, the Bayes posteriors are extraordinarily good, even though the data are not very informative. Figure 4 shows posterior distributions for the New York Mets using historical values for the prior parameter values. It can be seen that once the prior information is included the selected MLE produces an excellent posterior estimate.

3.5. Variance of the Estimates

When using a maximum likelihood estimate, $\hat{\theta}$, a common measure of variance is $-1/\ell''(\hat{\theta})$, where ℓ is the log-likelihood, $\ell = \log L$. In our situation, where L is expressed as a sum of component likelihoods (2.7), taking logs is not desirable. However, a few simple observations allow us to derive an approximation for the variance.

Since $\ell = \log L$, it follows that

$$\ell' = \frac{L'}{L} \qquad \text{and} \qquad \ell'' = \frac{LL'' - (L')^2}{L^2}. \tag{3.8}$$

We have $L'(\hat{\theta}) = 0$, $-\ell''(\hat{\theta}) = -L''(\hat{\theta}) \Big/ L(\hat{\theta})$, and using (2.7) yields the approximation

$$\begin{aligned}\mathrm{Var}(\hat{\theta} \mid \mathbf{y}) \approx -\left[\frac{L^{*''}(\hat{\theta} \mid \mathbf{y})}{L^*(\hat{\theta} \mid \mathbf{y})}\right]^{-1} &\approx \frac{\sum_{i=1}^{M} L^*(\hat{\theta}_i \mid \mathbf{y}, \mathbf{z}_i)}{\sum_{i=1}^{M} L^{*''}(\hat{\theta}_i \mid \mathbf{y}, \mathbf{z}_i)} \\ &= \frac{1}{n} \frac{\sum_{i=1}^{M} \hat{\theta}_i(1-\hat{\theta}_i) L^{*''}(\hat{\theta}_i \mid \mathbf{y}, \mathbf{z}_i)}{\sum_{i=1}^{M} L^{*''}(\hat{\theta}_i \mid \mathbf{y}, \mathbf{z}_i)},\end{aligned} \tag{3.9}$$

using (2.8) and (3.4). Of course, for the selected data, the variances are much higher than if we observed the entire data set. We thus modify (3.9) to account for the fact that we did not observe n Bernoulli trials, but only $m^* + 1$. Since our calculations are analogous to those in the EM algorithm, we could adjust (3.9) as in equation (3.26) of Dempster, Laird and Rubin (1977), where they show that the ratio of the complete-data variance to incomplete-data variance is given by the derivative of the EM mapping. However, in our case we have an even simpler answer, and assume $\mathrm{Var}(\hat{\theta} \mid \mathbf{y}, \mathbf{z}) \Big/ \mathrm{Var}(\hat{\theta} \mid \mathbf{y}) = (m^* + 1)/n$. Thus, we modify (3.9) to

$$\mathrm{Var}(\hat{\theta}\,|\,\mathbf{y}) \approx \frac{1}{m^*+1}\,\frac{\sum_{i=1}^{M} \hat{\theta}_i(1-\hat{\theta}_i)\,L^{*\prime\prime}(\hat{\theta}_i\,|\,\mathbf{y},\mathbf{z}_i)}{\sum_{i=1}^{M} L^{*\prime\prime}(\hat{\theta}_i\,|\,\mathbf{y},\mathbf{z}_i)}.$$

It turns out that this approximation works quite well, better than the "single" Taylor series approximation for the variance of the MLE of θ. However, the double Taylor series argument of Section 3.3 results in an improved approximation. Starting from the fact that $\ell'' = [LL''-(L')^2]/L^2$, we write

$$\begin{aligned}\mathrm{Var}(\hat{\theta}\,|\,\mathbf{y}) &\approx \frac{-n}{m^*+1}\left[\frac{L^{*\prime\prime}}{L^*}-\left(\frac{L^{*\prime}}{L^*}\right)^2\right]^{-1} \\ &= \frac{-n}{m^*+1}\left[\sum_{i,j}\frac{L^{*\prime\prime}(\hat{\theta}_j\,|\,\mathbf{y},\mathbf{z}_i)}{L^*(\hat{\theta}_j\,|\,\mathbf{y},\mathbf{z}_i)}-\left(\sum_{i,j}\frac{L^{*\prime}(\hat{\theta}_j\,|\,\mathbf{y},\mathbf{z}_i)}{L^*(\hat{\theta}_j\,|\,\mathbf{y},\mathbf{z}_i)}\right)^2\right]^{-1}. \qquad (3.10)\end{aligned}$$

Table 3 compares the approximation in (3.10) to values obtained by calculating $\ell''(\hat{\theta})$ exactly. It can be seen that the approximation is quite good for moderate to large m^*, and still is acceptable for small m^*.

⟨Table 3 about here⟩

We can also compute Bayesian variance estimates. Proceeding analogously to Section 3.4, the Bayesian variance would be an average of beta variances

$$\mathrm{Var}(\theta\,|\,\mathbf{y},a,b) \approx \frac{1}{M}\sum_{i=1}^{M}\frac{(k^*+S_{\mathbf{z}_i}+a)(n-k^*-S_{\mathbf{z}_i}+b)}{(n+a+b)^2(n+a+b+1)}. \qquad (3.11)$$

However, as before, we must adjust this variance to account for the fact that we only observe m^*+1 trials. We do this adjustment by replacing the term $(n+a+b+1)$ by $(m^*+1+a+b+1)$. The resulting estimate behaves quite reasonably, yielding estimates close to the exact MLE values for moderate m^* and $a = b = 1$. These values are also displayed in Table 3.

As expected, the standard error is sensitive to m^*, yielding large limits when m^* is small. Figures 5 and 6 show these limits for the 1992 season of Dave Winfield and the Mets, respectively. Although the estimates and standard errors are quite variable, note that the true batting average and winning percentage are always within one standard deviation of the selected MLE.

⟨Figures 5 and 6 about here⟩

4. Brownian Motion Approximation to the Likelihood

The last terms in the expressions (2.2) and (2.3),

$$\sum_{z \in \mathcal{Z}^*} \theta^{S_z}(1-\theta)^{m-S_z} = 1-\sum_{i=1}^{i^*} c_i \theta^{n_i+1-i}(1-\theta)^{i-1}, \tag{4.1}$$

are complicated to compute. But we can approximate these terms with functions derived from a consideration of Brownian motion. The resulting approximate likelihood can then be maximized to find an approximate MLE.

This was done for the data in Table 1. The second entry in each case is the approximate MLE. It can be seen that the approximate MLEs are excellent. In the 48 cases with $m \geq 60$, the approximate and exact MLEs never differ by more than .006. In fact, even for the smaller values of m, the exact and approximate MLEs are very close. In only three cases, all with $m = 5$, do the two differ by more than .01.

To develop our approximation, note from Appendix 1 that the expression in (4.1) is equal to $P_\theta\Big(S_j/(j+1) \leq r^* \text{ for } j = 1,\ldots,m\Big)$ where $Z_1, Z_2,\ldots$ are independent Bernoulli(θ) random variables and $S_j = \sum_{i=1}^{j} Z_i$. We rewrite the inequality $S_j/(j+1) \leq r^*$ as

$$S_j^* = \sum_{i=1}^{j} \frac{Z_i-\theta}{\sqrt{\theta(1-\theta)}} \leq \frac{(j+1)r^*-j\theta}{\sqrt{\theta(1-\theta)}} = b_\theta + \eta_\theta j\ ,$$

where $\eta_\theta = (r^*-\theta)\Big/\sqrt{\theta(1-\theta)}$ and $b_\theta = r^*\Big/\sqrt{\theta(1-\theta)}$. Now the vector $(S_1^*,\ldots,S_m^*)$ has the same means, variances and covariances as $\Big(W(1),\ldots,W(m)\Big)$, where W(t) is standard (mean 0 and variance 1) Brownian motion. So we can approximate

$$P_\theta\left(\frac{S_j}{j+1} \leq r^*,\ j = 1,\ldots,m\right) = P_\theta\Big(S_j^* \leq b_\theta + \eta_\theta j\,,\ j = 1,\ldots,m\Big)$$

by

$$P\Big(W(t) \leq b_\theta + \eta_\theta t, \quad 0 \leq t \leq m\Big).$$

If we define τ as the first passage time of W(t) through the linear boundary $b_\theta + \eta_\theta t$, that is, $\tau = \inf\{t : W(t) > b_\theta + \eta_\theta t\}$, then

$$\begin{aligned} P\Big(W(t) \leq b_\theta + \eta_\theta t, \quad 0 \leq t \leq m\Big) &= P(\tau > m) \\ &= 1-P(\tau \leq m) \\ &= \Phi\left(\frac{b_\theta}{\sqrt{m}} + \eta_\theta\sqrt{m}\right) - e^{-2b_\theta\eta_\theta}\,\Phi\left(-\frac{b_\theta}{\sqrt{m}} + \eta_\theta\sqrt{m}\right), \end{aligned}$$

where Φ is the standard normal cdf and the last equality is from (3.15) of Siegmund (1985).

Because S_j^* is a discrete process, the first time $S_j^* > b_\theta + \eta_\theta j$ it will in fact exceed the boundary by a positive amount. Also, W(t) may exceed $b_\theta + \eta_\theta t$ for some $0 \leq t \leq m$, even if $\Big(W(1),\ldots,W(m)\Big)$ does

not. So the probability we want is, in fact, larger than the approximation. Siegmund (1985, p. 50) suggests that this approximation will be improved if b_θ is replaced by $b_\theta + \rho$, where ρ is an appropriately chosen constant. By trial and error, we found that $\rho = .85$ produced good approximate MLEs. Thus, to obtain the approximate MLEs in Table 1, expression (4.1) was replaced by

$$\Phi\left(\frac{b_\theta + \rho}{\sqrt{m}} + \eta_\theta\sqrt{m}\right) - e^{-2(b_\theta + \rho)\eta_\theta}\,\Phi\left(-\frac{b_\theta + \rho}{\sqrt{m}} + \eta_\theta\sqrt{m}\right) \tag{4.2}$$

in (2.3) and the approximate likelihood was numerically maximized. (The zero of the derivative was found using a symbolic manipulation program, Maple V, just as the exact MLEs were found.)

5. Discussion

If we are told that Dave Winfield is 8-for-his-last-17, the somewhat unhappy conclusion is that there really isn't very much information being given. However, the somewhat surprising observation is that there is some information. Although we cannot hope to recover the complete data MLE with any degree of accuracy, we see in Figures 5 and 6 that ± 2 standard deviations of the selected data MLE always contains the complete data MLE. Indeed, in almost every case the complete data MLE is within one standard deviation of the selected data MLE. Moreover, the selected data estimates behave as expected. In particular, n (either the number of games or at-bats) increases, the ratio 8-for-17 looks worse, that is, it results in a smaller value of the selected MLE. This is as it should be, since, for a given success probability, longer strings (larger values of n) will produce larger maxima. Also, mainly due to the method of construction, the standard deviation of the selected MLE directly reflects the amount of information it contains, through the ratio m^*/n.

Baseball is a sport that is well-known for its accumulation of data. This readily translates into an enormous amount of prior information which can be used for estimation. In Figure 4 we saw how the New York Mets' prior information completely overwhelms the selected data (and produces very good estimates). This is, in fact, not an extreme case. If a picture similar to Figure 4 is constructed for Dave Winfield (with prior mean .285 and standard deviation .021), the resulting posterior is virtually a spike, no matter what data are used.

⟨Table 4 about here⟩

Throughout this paper we have assumed that the observed selected data consist of k^*, m^*, and n. However, typically the value of n is not reported, so the data are really only k^* and m^*. During the baseball season it is quite easy to estimate n, especially for ballplayers who play regularly. Moreover, once n reaches a moderate value, its value has very little effect on that of the selected data MLE. For example, for an everyday player, we expect $n \approx 100$ by May, so the value of n will have little effect on MLEs based on $m^* \leq 20$. This is evident in Table 1, the likelihood functions of Figure 3, and also in Table 4, which explores some limiting behavior of the MLE. Although we don't know the exact expression for the limit as $n \to \infty$, two points are evident. Besides the fact that the effect of n diminishes as n grows, it is clear that $r^* = 8/17$ and $r^* = 16/34$ have different limits. Thus, there is much more information contained in the pair (k^*, m^*) than in the single number r^*.

Lastly, we report the observation that a colleague (Chuck McCulloch) made when looking at Figure 1, an observation that may have interest for the baseball fan. When m^* and n are close together a number of things occur. First, the selected data and complete data MLE are close, and second, the selected data standard deviation is smallest. Thus, the selected data MLE is a very good estimator. However, McCulloch's observation is that when $\hat{\theta} \approx \hat{p}$ (the complete data MLE), which usually implies $m^* \approx n$, then a baseball player is in a batting slump (his current batting average is his maximum success ratio). This definition of a slump is based only on the players relative performance,

relative to his own "true" ability. A major drawback of our current notion of a slump is that it is usually based on some absolute measure of hitting ability, making it more likely that a .225 hitter, rather than a .300 hitter, would appear to be in a slump. (If a player is 1-for-his-last-10, is he in a slump? The answer depends on how good a hitter he actually is. For example, Tony Gwynn's slump could be Charlie O'Brian's hot streak!) If we examine Figure 1, Dave Winfield was in a bad slump during at-bats 156-187 (he was 5-for-31 = .161) and 360-377 (3-for-17 = .176), for in both cases his maximum hitting ability, r^*, is virtually equal to the MLEs. Similar observations can be made for Figure 2 and the Mets, particularly for games 45-70 (although many would say the New York Mets' entire 1992 season was a slump!). But the message is clear, you are in a slump if your complete data MLE is equal to your selected data MLE, for then your maximum hitting (or winning) ability is equal to your average ability.

Acknowledgements. We thank Steve Hirdt of the Elias Sports Bureau for providing us with detailed data for the 1992 Major League Baseball Season. We also thank Marty Wells for numerous conversations concerning the Gibbs/EM algorithm implementation.

Appendices: Technical Details

Appendix 1: Derivation of Combinatorial Formula for Likelihood

In this section we derive the exact expressions (2.2) and (2.3) for $L(\theta \mid \mathbf{y})$. The reported data is "$k^*$ successes in the last m^* trials". We have not specified exactly how this report was determined. In the baseball example, we do not know exactly how the announcer decided to report "k^* out of m^*". But what we have assumed is that the complete data $(\mathbf{y},\mathbf{z})$ consists of a vector $\mathbf{y}$, with k^* 1s and m^*+1-k^* 0s, and a vector $\mathbf{z} \in \mathcal{Z}^*$, a vector that satisfies (2.1). The likelihood is then

$$\sum_{\mathbf{y},\mathbf{z}} \theta^{S_{\mathbf{y},\mathbf{z}}}(1-\theta)^{n-S_{\mathbf{y},\mathbf{z}}},$$

where the sum is over all $(\mathbf{y},\mathbf{z})$ that give the reported data and $S_{\mathbf{y},\mathbf{z}} = \sum y_i + \sum z_i = k^* + S_{\mathbf{z}}$. We have not specified exactly what all the possible $\mathbf{y}$ vectors are, but, for each possible $\mathbf{y}$, $\mathbf{z}$ can be any element in $\mathcal{Z}^*$. Thus if C is the number of possible $\mathbf{y}$ vectors then the likelihood is

$$C\,\theta^{k^*}(1-\theta)^{m^*+1-k^*} \sum_{\mathbf{z} \in \mathcal{Z}^*} \theta^{S_{\mathbf{z}}}(1-\theta)^{m-S_{\mathbf{z}}}. \tag{A.1}$$

Dropping the constant C, which is unimportant for likelihood analysis, yields (2.2).

Let $\mathcal{Z}$ denote the set of all sequences $(z_1,\ldots,z_m)$ of length m of 0s and 1s. Then, the sum in (A.1) is

$$\begin{aligned} \sum_{\mathbf{z} \in \mathcal{Z}^*} \theta^{S_{\mathbf{z}}}(1-\theta)^{m-S_{\mathbf{z}}} &= \sum_{\mathbf{z} \in \mathcal{Z}} \theta^{S_{\mathbf{z}}}(1-\theta)^{m-S_{\mathbf{z}}} - \sum_{\mathbf{z} \in \mathcal{Z}^{*c}} \theta^{S_{\mathbf{z}}}(1-\theta)^{m-S_{\mathbf{z}}} \\ &= 1 - \sum_{\mathbf{z} \in \mathcal{Z}^{*c}} \theta^{S_{\mathbf{z}}}(1-\theta)^{m-S_{\mathbf{z}}}. \end{aligned} \tag{A.2}$$

This sum over $\mathcal{Z}^{*c}$ is the sum that appears in (2.3), as we now explain.

The set $\mathcal{Z}^*$ is the set of all $\mathbf{z}$s that satisfy (2.1). Let $S_j = \sum_{i=1}^{j} z_i$. Then $\mathcal{Z}^{*c}$ is the set of all $\mathbf{z}$s that satisfy

$$\frac{k^* + S_j}{m^* + 1 + j} = \frac{k^* + \sum_{i=1}^{j} z_i}{m^* + 1 + j} > r^* = \frac{k^*}{m^*}, \qquad \text{for some } j = 1,\ldots,m.$$

But, $(k^* + S_j)/(m^* + 1 + j) > k^*/m^*$ if and only if $S_j/(j+1) > r^*$. So, $\mathcal{Z}^{*c}$ is the set of all $\mathbf{z}$s that satisfy

$$\frac{S_j}{j+1} > r^*, \qquad \text{for some } j = 1,\ldots,m.$$

Now, to complete our derivation of (2.3) we must show that the sums in (A.2) and (2.3) are equal; that is,

$$\sum_{\mathbf{z} \in \mathcal{Z}^{*c}} \theta^{S_{\mathbf{z}}}(1-\theta)^{m-S_{\mathbf{z}}} = \sum_{i=1}^{i^*} c_i\, \theta^{n_i+1-i}(1-\theta)^{i-1}. \tag{A.3}$$

To show this we must explain what the constants i^*, n_i and c_i are.

First the value of i^*-1 is the maximum number of 0s that can occur in $(z_1,\ldots,z_j)$ if $S_j/(j+1) > r^*$. This is because $S_j/(j+1) > r^* \Leftrightarrow S_j > r^*(j+1) \Leftrightarrow j - S_j < j - r^*(j+1)$, and hence

$$\begin{aligned} j - S_j < j - r^*(j+1) &= j(1-r^*) - r^* \le m(1-r^*) - r^* \\ &= (n-m^*-1)(1-r^*) - r^* \\ &= (n-m^*)(1-r^*) - 1 \,. \end{aligned}$$

Next, suppose that when $S_j/(j+1)$ first exceeds r^*, the number of 0s in $(z_1,\ldots,z_j)$ is $j-S_j = j'-1$. Then this must happen on trial $j = n_{j'}$, because if on trial $n_{j'}$, $n_{j'} - S_{n_{j'}} = j'-1$, then

$$\frac{S_{n_{j'}}}{n_{j'}+1} = \frac{n_{j'}-j'+1}{n_{j'}+1} = 1 - \frac{j'}{n_{j'}+1} > 1 - \frac{j'}{\left(\frac{j'}{1-r^*} - 1\right) + 1} = r^* \,.$$

But if $j - S_j = j'-1$ and $j < n_{j'}$, then

$$\frac{S_j}{j'+1} = \frac{j-j'+1}{j+1} = 1 - \frac{j'}{j+1} \le 1 - \frac{j'}{\left(\frac{j'}{1-r^*} - 1\right) + 1} = r^* \,.$$

To compute the sum in (A.3), we partition $\mathcal{Z}^{*c}$ into sets $\mathcal{Z}_0,\ldots,\mathcal{Z}_{i^*-1}$ where $\mathcal{Z}_i$ is the set of $\mathbf{z}$s such that $(z_1,\ldots,z_j)$ contains exactly i 0s if $S_j/(j+1)$ is the first term to exceed r^*. That is,

$$\mathcal{Z}_i = \left\{ \mathbf{z}\colon j - S_j = i \text{ at the } j \text{ where } S_j/(j+1) > r^* \text{ and } S_{j'}/(j'+1) \le r^* \text{ for } 1 \le j' < j \right\}.$$

If $\mathbf{z} \in \mathcal{Z}_i$, then in fact $j = n_{i+1}$ from our argument above. Let $(z_1,\ldots,z_{n_{i+1}})$ be a sequence such that $S_{n_{i+1}}/(n_{i+1}+1) > r^*$ and $S_{j'}/(j'+1) \le r^*$ for $1 \le j' < n_{i+1}$. (So the vector $(z_1,\ldots,z_{n_{i+1}})$ contains i 0s and $n_{i+1}-i$ 1s.) This initial sequence can be completed in any way to produce a $\mathbf{z} \in \mathcal{Z}_i$. The sum of $\theta^{S_\mathbf{z}}(1-\theta)^{m-S_\mathbf{z}}$ over all $\mathbf{z}$s with this initial sequence is $\theta^{n_{i+1}-i}(1-\theta)^i$, because the sum over all the parts that could be added to this initial sequence is 1. Note that we get the same value $\theta^{n_{i+1}-i}(1-\theta)^i$, regardless of which initial sequence we choose. So if c_{i+1} is the number of different initial sequences that could form $\mathbf{z}$s in $\mathcal{Z}_i$, then

$$\begin{aligned} \sum_{\mathbf{z} \in \mathcal{Z}^{*c}} \theta^{S_\mathbf{z}}(1-\theta)^{m-S_\mathbf{z}} &= \sum_{i'=0}^{i^*-1} \sum_{\mathbf{z} \in \mathcal{Z}_{i'}} \theta^{S_\mathbf{z}}(1-\theta)^{m-S_\mathbf{z}} \\ &= \sum_{i=1}^{i^*} c_i \, \theta^{n_i+1-i}(1-\theta)^{i-1} \,, \end{aligned}$$

which is equation (A.3). It only remains to verify that the formula in Section 2.1 is the correct formula for c_i. The value of c_1 is the number of initial sequences with $1-1 = 0$ 0s. Of course, $c_1 = 1$, as defined. Suppose $c_1,\ldots,c_i$ are correctly defined. Then we will show that the formula,

$$c_{i+1} = \binom{n_{i+1}-1}{i} - \sum_{j=1}^{i} c_j \binom{n_i+1-1-n_j}{i+1-j}, \tag{A.4}$$

from Section 2.1 is correct. The value of $\binom{n_{i+1}-1}{i}$ is the number of all sequences $(z_1,\ldots,z_{n_{i+1}})$ that end in 1 and have exactly i 0s. From this we must subtract those sequences for which $S_j/(j+1) > r^*$ for some $j < n_{i+1}$. If $S_j/(j+1) > r^*$ for the first time at j' and if $j' - S_{j'} = i'$, then j' must equal $n_{i'+1}$. Among all sequences $(z_1,\ldots,z_{n_{i+1}})$ that end in 1 and have exactly i 0s, there are

$c_{i'+1}\binom{n_{i+1}-1-n_{i'+1}}{i-i'}$ that first exceed r^* at $n_{i'+1}$ with i' 0s. The value of $c_{i'+1}$ is the number of initial sequences $(z_1,\ldots,z_{n_{i'+1}})$ and the combinatorial term is the number of sequences $(z_{n_{i'+1}+1},\ldots,z_{n_{i+1}-1})$ containing the remaining $i-i'$ 0s. Summing these terms for $i' = 0,\ldots,i-1$, changing the summation index to $j = i'+1$, yields the sum in (A.4). Thus the formula for $c_1,\ldots,c_{i^*}$ is correct.

Appendix 2: Calculating Likelihoods with Gibbs Sampling

Gibbs sampling calculations for likelihood functions is actually a mixture of some EM algorithm ideas (Dempster, Laird and Rubin, 1977) and an implementation of successive substitution sampling (Gelfand and Smith, 1990).

As in the EM algorithm, we start with $L(\theta \mid \mathbf{y})$ as the likelihood of interest, based on the "incomplete" (but observed) data $\mathbf{y}$. The augmented data is denoted by $\mathbf{z}$, yielding the complete data likelihood $L(\theta \mid \mathbf{y}, \mathbf{z})$ which satisfies

$$L(\theta \mid \mathbf{y}) = \sum_{\mathcal{Z}^*} L(\theta \mid \mathbf{y}, \mathbf{z}) \, . \tag{A.5}$$

The set $\mathcal{Z}^*$ may be quite complicated, taking into account all the restrictions imposed on the incomplete data likelihood. However, it is often the case that we will be able to sample from this set.

Now normalize both likelihoods (as in Section 2.2) to $L^*(\theta \mid \mathbf{y})$ and $L^*(\theta \mid \mathbf{y}, \mathbf{z})$, and define $k(\mathbf{z} \mid \mathbf{y}, \theta)$ as in (2.6). Then,

$$L^*(\theta \mid \mathbf{y}) = \int_\Theta \Big(\sum_{\mathcal{Z}^*} L^*(\theta \mid \mathbf{y}, \mathbf{z}) \; k(\mathbf{z} \mid \mathbf{y}, \theta') \Big) \, L^*(\theta' \mid \mathbf{y}) \, d\theta' \, . \tag{A.6}$$

To verify equation (A.6), write

$$\int_\Theta \Big(\sum_{\mathcal{Z}^*} L^*(\theta \mid \mathbf{y}, \mathbf{z}) \; k(\mathbf{z} \mid \mathbf{y}, \theta') \; L^*(\theta' \mid \mathbf{y}) \Big) d\theta' = \sum_{\mathcal{Z}^*} L^*(\theta \mid \mathbf{y}, \mathbf{z}) \Big(\int_\Theta L^*(\theta' \mid \mathbf{y}, \mathbf{z}) \, d\theta' \Big), \tag{A.7}$$

by interchanging the order of the sum and integral and noting that $k(\mathbf{z} \mid \mathbf{y}, \theta') \; L^*(\theta' \mid \mathbf{y}) = L^*(\theta' \mid \mathbf{y}, \mathbf{z})$. Now the integral on the RHS of (A.7) is equal to 1, and the remaining sum is just (A.5). Thus, from equation (A.6) and the results of Gelfand and Smith (1990), we can calculate $L^*(\theta \mid \mathbf{y})$ by successively sampling from $L^*(\theta \mid \mathbf{y}, \mathbf{z})$ and $k(\mathbf{z} \mid \mathbf{y}, \theta)$.

Note that the implementation of the Gibbs sampler is a totally frequentist implementation. It only relies on the finiteness of the integral of $\int_\Theta L(\theta \mid \mathbf{y}) \, d\theta$. We can, however, interpret the finiteness of this integral as using a flat prior for θ, that is, $\pi(\theta) = 1$. With the additional "parameter" $\mathbf{z}$, we then have the two full posterior distributions $\pi(\theta \mid \mathbf{y}, \mathbf{z}) \Big(= L^*(\theta \mid \mathbf{y}, \mathbf{z}) \Big)$ and $\pi(\mathbf{z} \mid \mathbf{y}, \theta) \Big(= k(\mathbf{z} \mid \mathbf{y}, \theta) \Big)$. Iteratively, sampling from these densities yields a sample from the marginal posterior $\pi(\theta \mid \mathbf{y})$ $\Big(= L^*(\theta \mid \mathbf{y}) \Big)$.

References

Baseball Encyclopedia, Seventh Edition (1988). Joseph L. Reichler, Editor. New York: Macmillan Publishing Company.

Bayarri, M.J. and DeGroot, M. (1986a). Bayesian analysis of selection models. Technical Report 365, Dept. of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania.

Bayarri, M.J. and DeGroot, M. (1986b). Information in selection models. Technical Report 368, Dept. of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania.

Bayarri, M.J. and DeGroot, M. (1991). The analysis of published significant results. Technical Report 91-21, Dept. of Statistics, Purdue University, West Lafayette, Indiana.

Carlin, B.P. and Gelfand, A.E. (1992). Parameter likelihood inference for record breaking problems. Technical Report, Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota.

Cleary, R.J. (1993). Models for selection bias in meta-analysis. Ph.D. Thesis, Biometrics Unit, Cornell University, Ithaca, New York.

Dawid, A.P. and Dickey, J.M. (1977). Likelihood and Bayesian inference from selectively reported data. *J. Amer. Statist. Assoc.* **72**, 845-850.

Dear, K.B.G. and Begg, C.B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statist. Sci.* **7**, 237-245.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-37.

Gelfand, A.E. and Carlin, B.P. (1991). Maximum likelihood estimation for constrained or missing data models. Technical Report, Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Soc.* **85**, 398-409.

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54**, 657-699.

Hedges, L.V. (1992). Modeling publication selection effects in meta-analysis. *Statist. Sci.* **7**, 246-255.

Iyengar, S. and Greenhouse, J.B. (1988). Selection models and the file drawer problem. *Statist. Sci.* **3**, 109-135.

Mosteller, F. and Chalmers, T.C. (1992). Some progress and problems in meta-analysis of clinical trials. *Statist. Sci.* **7**, 227-236.

Rosenthal, R. (1979). The "file drawer" problem and tolerance for null results. *Psychol. Bull.* **86**, 638-641.

Siegmund, D. (1985). *Sequential Analysis.* New York: Springer-Verlag.

Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via a Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser. B* **55**, 3-24.

Table 1. Exact and approximate MLEs calculated from the exact likelihood (2.3) and the approximation described in (4.2). The first entry is the exact MLE and the second entry is the approximate MLE.

		m*							
r*	m	5		25		45		200	
	5	0.109	0.096	0.170	0.167	0.183	0.180	0.196	0.196
	20	0.088	0.079	0.150	0.145	0.167	0.164	0.191	0.190
1/5	60	0.084	0.078	0.139	0.134	0.156	0.152	0.185	0.184
	100	0.084	0.078	0.136	0.132	0.152	0.149	0.182	0.181
	200	0.084	0.078	0.135	0.131	0.149	0.147	0.178	0.177
	5	0.261	0.263	0.359	0.360	0.376	0.377	0.394	0.394
	20	0.232	0.232	0.332	0.332	0.356	0.356	0.389	0.389
2/5	60	0.227	0.229	0.317	0.316	0.341	0.341	0.381	0.381
	100	0.227	0.228	0.314	0.313	0.336	0.336	0.377	0.377
	200	0.227	0.228	0.312	0.311	0.333	0.333	0.372	0.372
	5	0.437	0.448	0.553	0.559	0.572	0.577	0.593	0.594
	20	0.408	0.412	0.527	0.531	0.554	0.556	0.588	0.589
3/5	60	0.403	0.407	0.511	0.513	0.538	0.541	0.580	0.581
	100	0.403	0.407	0.507	0.510	0.534	0.535	0.577	0.578
	200	0.403	0.407	0.506	0.508	0.530	0.531	0.571	0.572
	5	0.632	0.644	0.754	0.763	0.773	0.779	0.794	0.795
	20	0.612	0.615	0.734	0.742	0.759	0.765	0.789	0.791
4/5	60	0.609	0.609	0.721	0.726	0.746	0.750	0.783	0.785
	100	0.609	0.609	0.718	0.722	0.742	0.745	0.780	0.782
	200	0.609	0.609	0.717	0.720	0.739	0.742	0.776	0.777

Table 2. Comparison of combinatoric MLE (obtained by differentiating the likelihood (2.3)), the MLE from the EM algorithm, the Gibbs/likelihood approximation of (3.3), the Bayes posterior mean using a beta(1,1) prior (the mean likelihood estimate), the Brownian motion-based approximation, and $\hat{p}$ (the complete data MLE). The at-bats were chosen from Dave Winfield's 1992 season.

At-bat	k^*	m^*	r^*	$\hat{p}$	Combinatoric MLE	EM MLE	Gibbs' Approx. MLE	Bayes mean	Brownian MLE
187	55	187	.294	.294	.294	.294	.294	.296	.294
188	4	11	.364	.298	.241	.240	.232	.237	.241
189	5	12	.417	.302	.292	.289	.276	.278	.293
190	5	13	.385	.300	.267	.267	.253	.258	.268
191	5	11	.455	.304	.321	.322	.303	.299	.324
339	12	39	.308	.298	.241	.240	.238	.225	.241
340	47	155	.303	.297	.273	.273	.273	.272	.273
341	13	41	.317	.299	.251	.251	.250	.243	.251
342	13	42	.310	.298	.245	.244	.242	.238	.245
343	14	43	.326	.300	.260	.261	.258	.252	.260
344	14	44	.318	.299	.254	.254	.246	.247	.254
345	4	11	.367	.301	.241	.240	.222	.234	.241
346	5	11	.456	.303	.321	.313	.300	.302	.325

Table 3. Comparison of standard deviations based on exact differentiation of the log likelihood, the Gibbs/likelihood approximation of (3.10), and the Bayes posterior standard deviation using a beta(1, 1) prior. The at-bats were chosen from Dave Winfield's 1992 season.

					Standard deviation		
At-bat	k^*	m^*	$\hat{p}$	MLE	Exact	Gibbs/Approx.	Bayes
187	55	187	.294	.294	.033	.033	.033
188	4	11	.298	.241	.083	.121	.117
189	5	12	.302	.292	.086	.122	.118
190	5	13	.300	.267	.080	.115	.112
191	5	11	.304	.321	.093	.129	.125
339	12	39	.298	.241	.046	.067	.065
340	47	155	.297	.273	.029	.036	.035
341	13	41	.299	.251	.046	.067	.065
342	13	42	.298	.245	.045	.065	.064
343	14	43	.300	.260	.046	.066	.064
344	14	44	.299	.254	.045	.064	.063
345	4	11	.301	.241	.083	.119	.116
346	5	11	.303	.321	.093	.131	.126

Table 4. Limiting behavior of the MLE for fixed $r^* = .471$, as $n \to \infty$.

	k^*/m^*			
n	8/17	16/34	64/136	128/272
25	.395	—	—	—
50	.368	.418	—	—
75	.360	.401	—	—
100	.359	.396	—	—
150	.358	.392	.456	—
200	.358	.391	.442	—
300	.358	.391	.435	.460
400	.357	.390	.431	.451
500	.357	.390	.431	.447
1000	.357	.390	.429	.442

Figure 1. The 1992 hitting record of Dave Winfield. The dashed line is the complete data MLE $\hat{p}$ = ratio of hits to at-bats, the solid line is r^*, the selected maximum ratio, and the dotted line is the selected data MLE, $\hat{\theta}$.

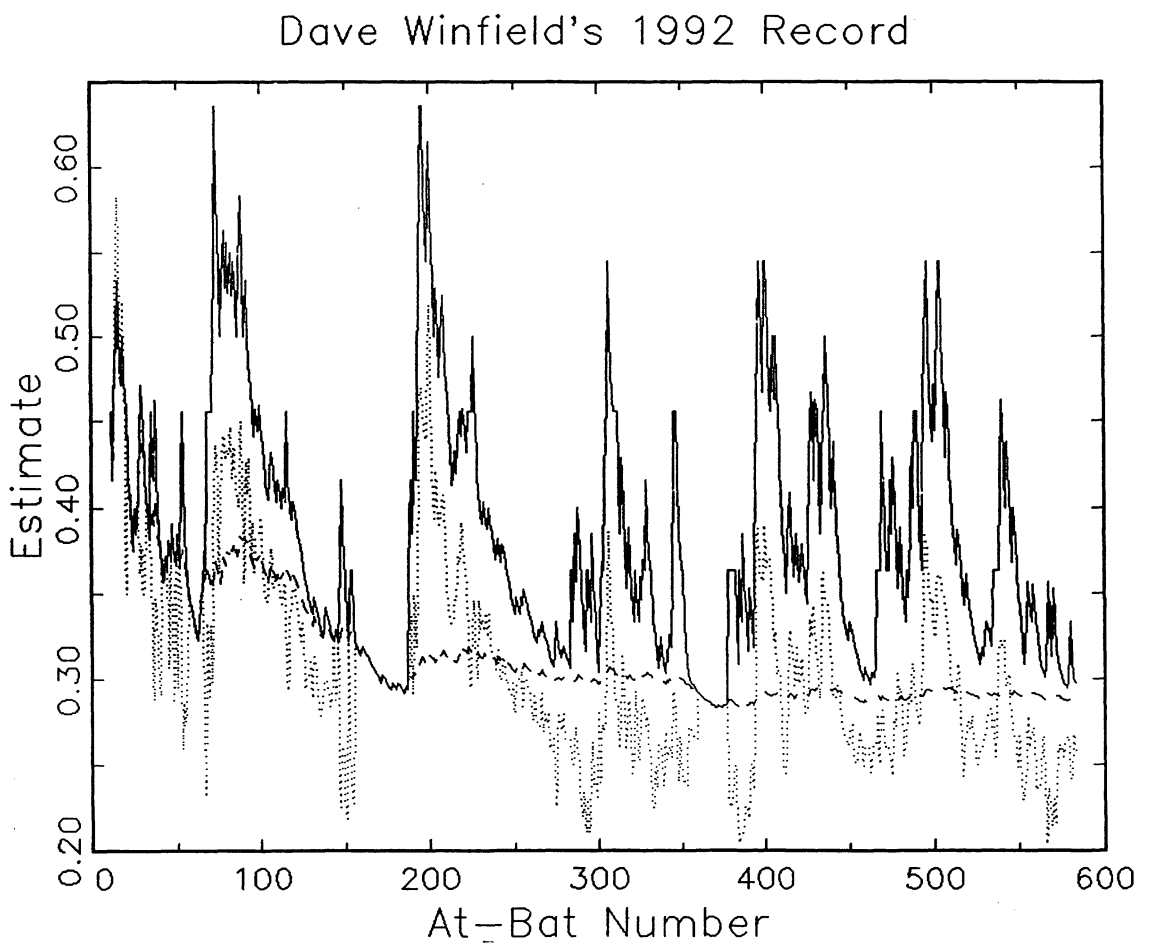

Figure 2. The 1992 won-loss record of the New York Mets. The dashed line is the complete data MLE $\hat{p}$ = ratio of wins/games played, the solid line is r^*, the selected maximum ratio, and the dotted line is the selected data MLE, $\hat{\theta}$.

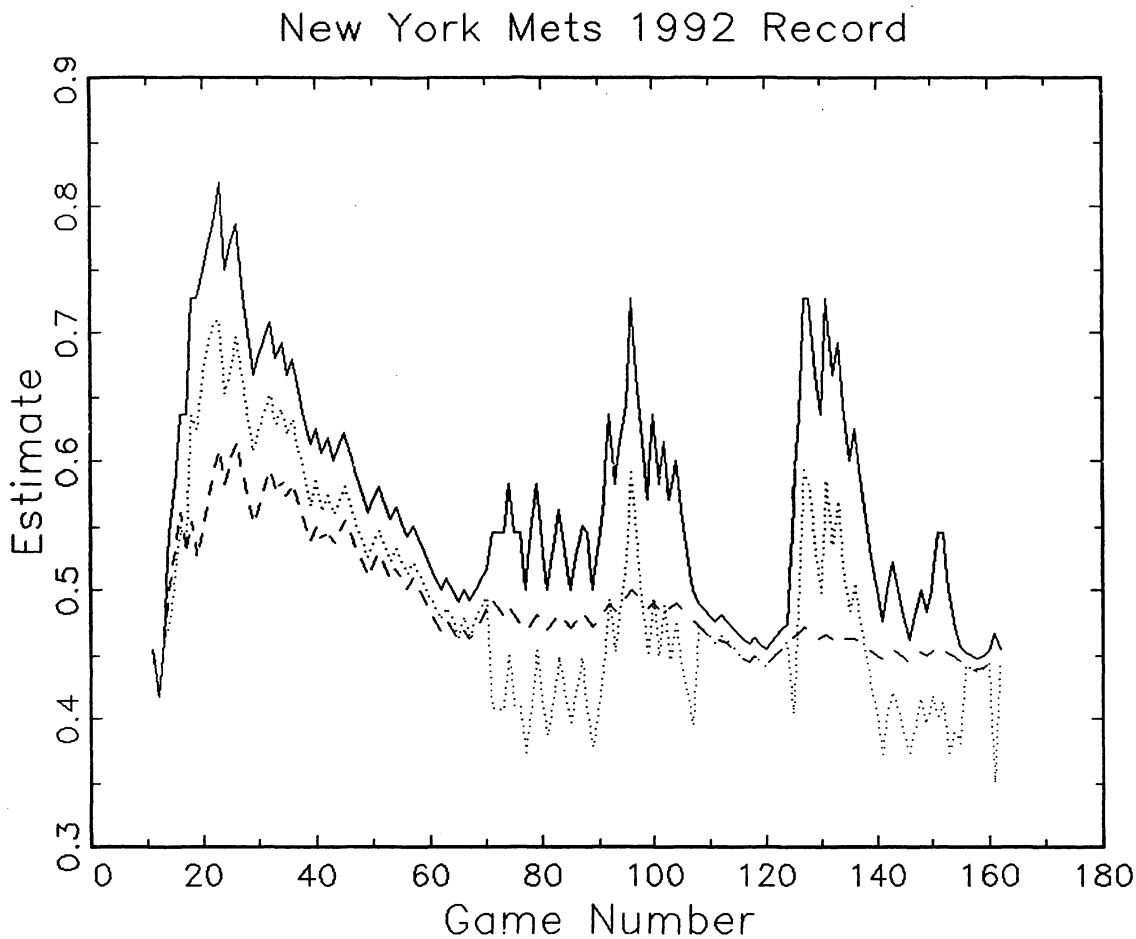

Figure 3. Likelihood functions for $k^* = 8$ and $m^* = 17$, for $n = 18$, 25, 50 and 100, normalized to have area = 1. The value $n = 18$ is the "naive" likelihood, and is represented by the dashed lines. The other three likelihoods are represented by solid lines. As n increases, both the modes and variances decrease.

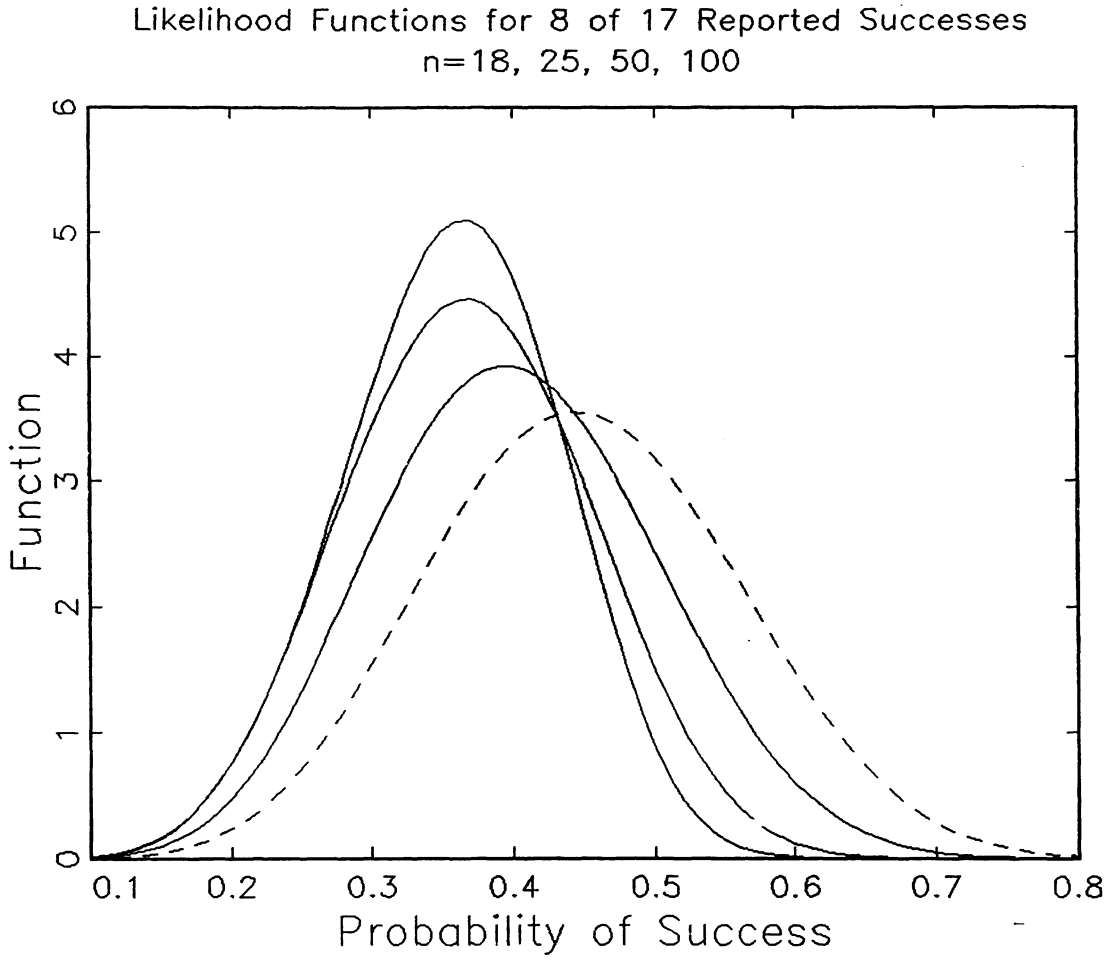

Figure 4. Posterior distributions for the New York Mets. The value $r^* = 8/11 = .727$ was actually achieved at games 19 and 131. The solid lines are posterior distributions based on beta priors with parameters $a = 9.951$ and $b = 11.967$, representing a mean of .454 and a standard deviation of .107, which are the Mets' overall past parameters. The two solid lines are based on $n = 19$ and 131 observations, with modes decreasing in n. The dotted lines are posteriors using a beta(1,1) prior with $n = 19$ and 131, hence are likelihood functions, again with modes decreasing in n. The dotted line is the naive likelihood, which assumes $n = 12$.

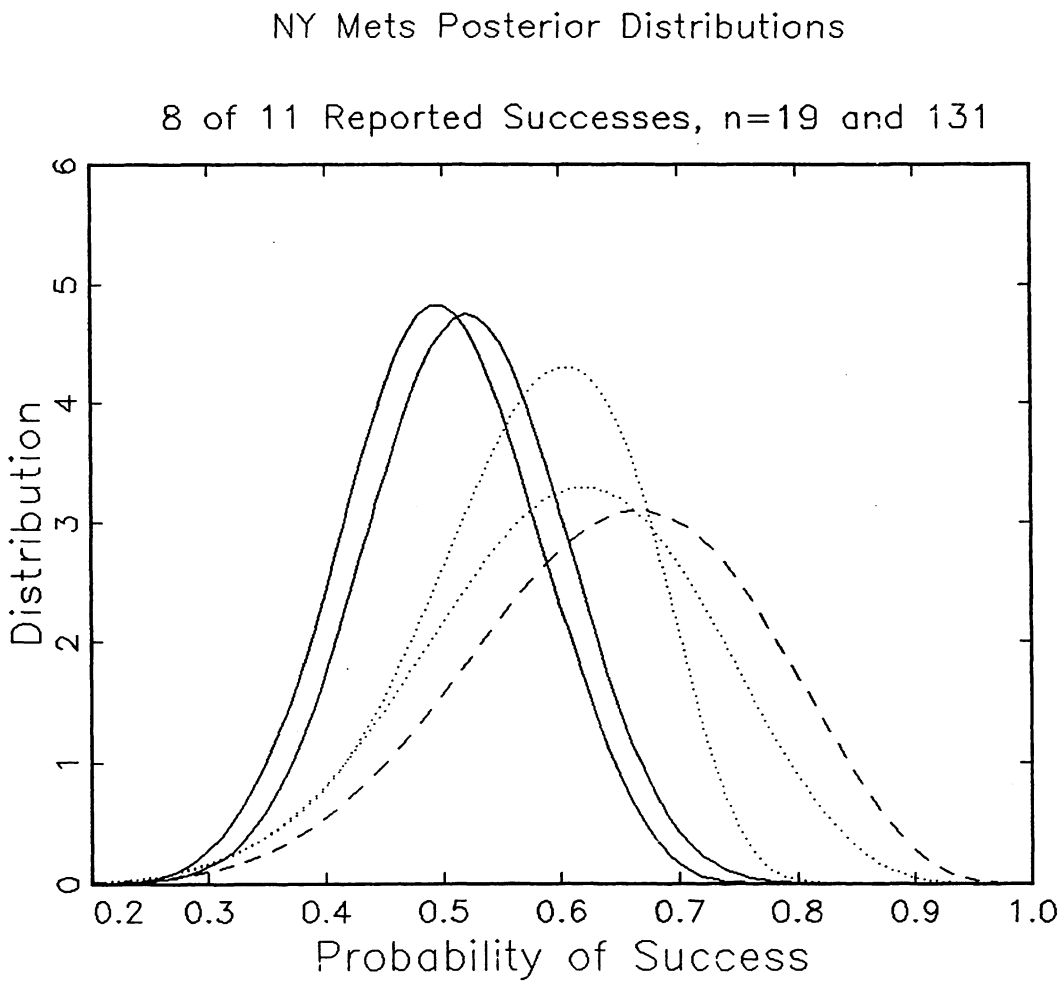

Figure 5. The 1992 hitting record of Dave Winfield. The dashed line is $\hat{p}$, the complete data MLE, and the solid line is $\hat{\theta}$, the selected data MLE. The standard deviation limits (dotted lines) are based on the selected data MLE.

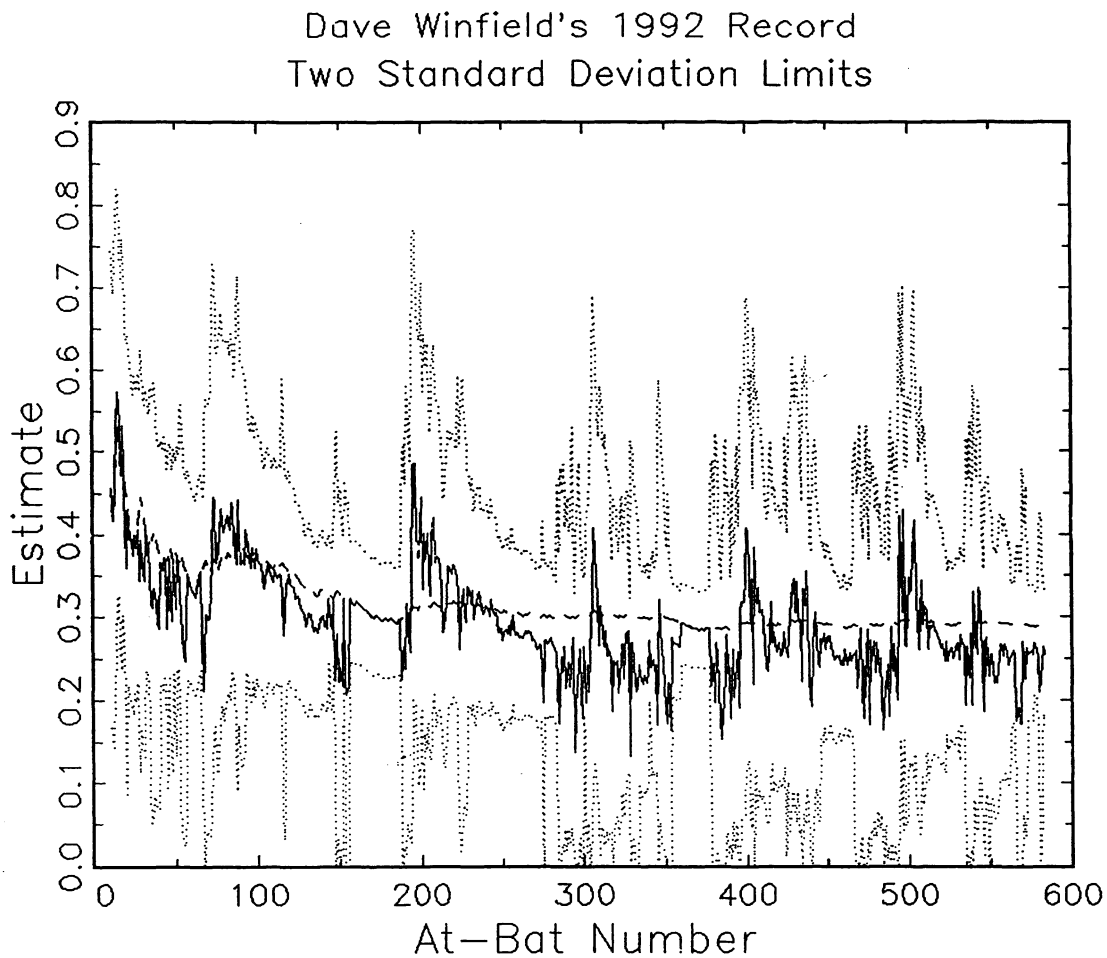

Figure 6. The 1992 won-loss record of the New York Mets. The dashed line is $\hat{p}$, the complete data MLE, and the solid line is $\hat{\theta}$, the selected data MLE. The standard deviation limits (dotted lines) are based on the selected data MLE.